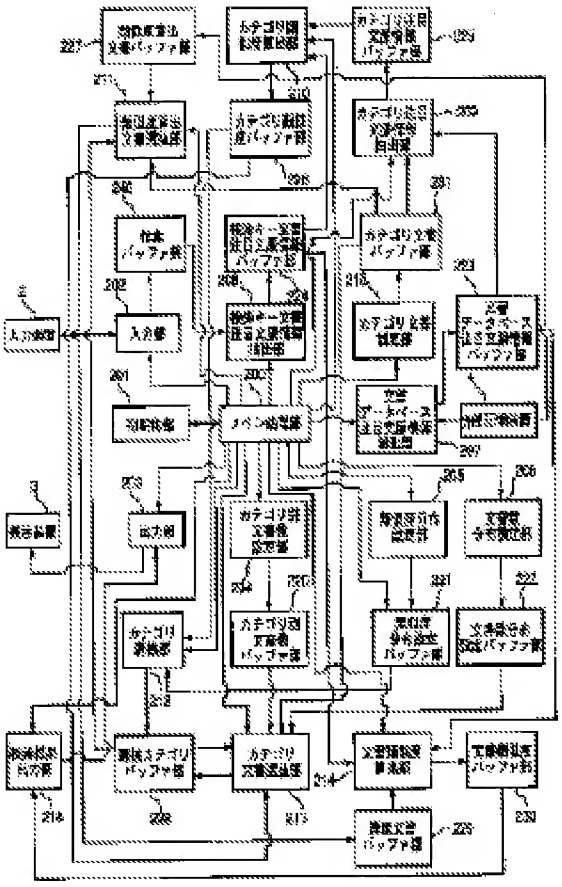
SIMILAR DOCUMENT RETRIEVING DEVICE AND SIMILAR DOCUMENT RETRIEVING METHOD

Publication number: JP10289246
Publication date: 1998-10-27
Inventor: TANOSAKI YASUO; NAKAMOTO YUKIO; NISHINA TAKUYA; KUBOTA NAOHIDE
Applicant: TOKYO SHIBAURA ELECTRIC CO; TOSHIBA COMPUTER ENG
Classification:
- international: ***G06F17/30;*** **G06F17/30;** (IPC1-7): G06F17/30
- european:
Application number: JP19970097630 19970415
Priority number(s): JP19970097630 19970415

Abstract of **JP10289246**

PROBLEM TO BE SOLVED: To provide a similar document retrieving device which efficiently retrieves document that is similar to a retrieval key document even when there are many documents. SOLUTION: This device inputs a retrieval key document from an input device 2, extracts document remarked context information that suggests the contents from each document in an external storage device 4 and the retrieval key document, extracts category remarked context information with a document group of different contents in the device 4 all together and calculates category similarity between the document remarked context information and the category remarked context information that is extracted from the document group of different contents in a lump. It selects a document group that calculates document similarity with the retrieval key document from a document database in accordance with the calculated category similarity and outputs identification information of a document that is retrieved based on each document similarity which is calculated by the document similarity calculating means as a retrieval result to a display device 3.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－289246

(43)公開日　平成10年(1998)10月27日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｃ０６Ｆ　15/403	３５０Ｃ
		15/40	３７０Ａ

審査請求　未請求　請求項の数６　ＯＬ　（全 13 頁）

(21)出願番号　特願平9－97630

(22)出願日　平成９年(1997)４月15日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(71)出願人　000221052
東芝コンピュータエンジニアリング株式会社
東京都青梅市新町３丁目３番地の１

(72)発明者　田野崎　康雄
東京都青梅市末広町二丁目九番地　株式会社東芝青梅工場内

(74)代理人　弁理士　三澤　正義

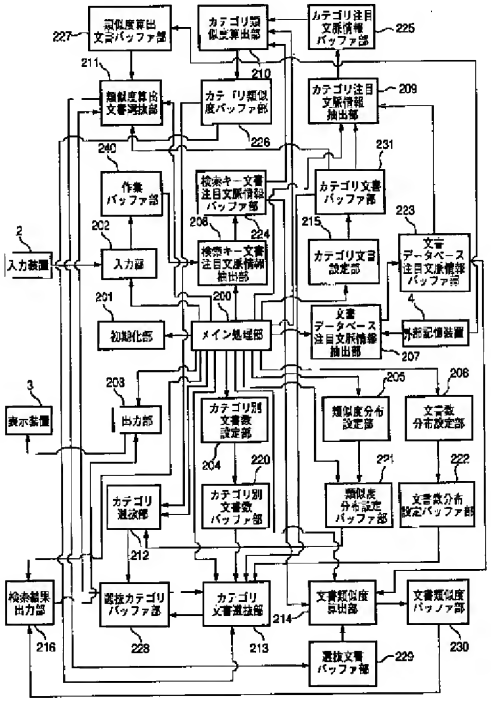
最終頁に続く

(54)【発明の名称】　類似文書検索装置及び類似文書検索方法

(57)【要約】

【課題】　本発明は、文書数が多い場合であっても、効率良く検索キー文書に類似している文書を検索することが可能な類似文書検索装置を提供する。

【解決手段】　入力装置２から検索キー文書を入し、外部記憶装置４中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、外部記憶装置４中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜し、この文書類似度算出手段により算出した各文書類似度を基に検索した文書の識別情報を検索結果として表示装置３へ出力するようにしたものである。



【特許請求の範囲】
【請求項１】　一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、
検索キー文書を入力する入力手段と、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報情報抽出手段と、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、
このカテゴリ類似度算出手段により算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜する類似度算出文書選抜手段と、
この類似度算出文書選抜手段より選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出する文書類似度算出手段と、
この文書類似度算出手段により算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力する出力手段と、
を有することを特徴とする類似文書検索装置。
【請求項２】　一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、
検索キー文書を作成し、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、
算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜し、
選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出し、
算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力すること、
を特徴とする類似文書検索方法。
【請求項３】　一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、
検索キー文書を入力する入力手段と、
検索対象文書のカテゴリ類似度分布を設定するカテゴリ類似度分布設定手段と、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報情報抽出手段と、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、
このカテゴリ類似度算出手段により算出したカテゴリ類似度とカテゴリ類似度分布設定手段により設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜する類似度算出文書選抜手段と、
この類似度算出文書選抜手段より選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出する文書類似度算出手段と、
この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力する出力手段と、
を有することを特徴とする類似文書検索装置。
【請求項４】　一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、
検索キー文書を作成し、
検索対象文書のカテゴリ類似度分布を設定し、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、
算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜し、
選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出し、
算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力すること、
を特徴とする類似文書検索方法。
【請求項５】　一文書を検索キーとして、文書群が内容

別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、
検索キー文書を入力する入力手段と、
検索対象文書のカテゴリ類似度分布を設定するカテゴリ類似度分布設定手段と、
検索対象文書の文書数分布を設定する文書数分布設定手段と、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報情報抽出手段と、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、
このカテゴリ類似度算出手段により算出したカテゴリ類似度とカテゴリ類似度分布設定手段により設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、前記文書数分布設定手段により設定した検索対象文書の文書数分布に応じた数の文書群に絞る類似度算出文書選抜手段と、
この類似度算出文書選抜手段より絞った文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に各文書類似度を算出する文書類似度算出手段と、
この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力する出力手段と、
を有することを特徴とする類似文書検索装置。
【請求項６】　一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、
検索キー文書を作成し、
検索対象文書のカテゴリ類似度分布を設定し、
検索対象文書の文書数分布を設定し、
文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、
文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、
検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、
算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、設定した検索対象文書の文書数分布に応じた数の文書群に絞り、
絞った文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に各文書類似度を算出し、
この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力すること、
を特徴とする類似文書検索方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、類似文書検索装置及び類似文書検索方法に関する。
【０００２】
【従来の技術】従来、文書を検索キーとして、その文書の内容に類似している文書を検索対象文書データベースから抽出する類似文書検索装置が提案されている。この類似文書検索装置は、検索キーとする文書中に含まれている単語と、検索対象文書データベースに格納されている各文書中に含まれている単語とを比較し、検索キーとする文書と検索対象文書データベースに格納されている各文書との類似度を算出し、その類似度の高低に応じて類似文書の抽出を行っている。
【０００３】このような類似度の算出方法としては、検索キーとする文書と、検索対象文書データベースに格納されている各文書に含まれている単語の種類や出現回数、出現場所等とから空間ベクトル法を使用して算出する方法が採用されている。
【０００４】
【発明が解決しようとする課題】しかしながら、上述した従来技術においては、検索対象文書データベースに格納されている各文書と、検索キーとする文書との類似度算出を、検索対象文書データベースに格納されている文書数分行うため、検索対象文書データベースに格納されている文書数が多い場合には、必然的に検索処理時間が増加するという課題があった。
【０００５】本発明は上記の課題を解決するためになされたものであり、検索対象文書データベースに格納されている文書数が多い場合であっても、検索処理時間を短縮でき、効率良く検索キーに類似している文書を検索することが可能な類似文書検索装置及び類似文書検索方法を提供することを目的とする。
【０００６】
【課題を解決するための手段】請求項１記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、検索キー文書を入力する入力手段と、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報情報抽出手段と、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏ま

りの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、このカテゴリ類似度算出手段により算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜する類似度算出文書選抜手段と、この類似度算出文書選抜手段より選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出する文書類似度算出手段と、この文書類似度算出手段により算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力する出力手段とを有することを特徴とするものである。

【０００７】請求項２記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、検索キー文書を作成し、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜し、選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出し、算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力することを特徴とするものである。

【０００８】請求項１記載の発明に係る類似文書検索装置の構成を使用した請求項２記載の発明に係る類似文書検索方法は、入力手段により、検索キー文書を作成し、文書注目文脈情報抽出手段により文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、カテゴリ注目文脈情報情報抽出手段により文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出する。

【０００９】さらに、カテゴリ類似度算出手段により、検索キー文書から抽出した文書注目文脈情報と内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、類似度算出文書選抜手段により算出したカテゴリ類似度に応じて前記文書データベース中から検索キー文書との間で文書類似度を算出する文書群を選抜し、出力手段により算出した各文書類似度を基に検索した文書の識別情報を検索結果として出力するものである。

【００１０】これにより、文書データベースに格納されている文書数の増大とともに、検索処理時間が増加するという従来の課題を解決し、文書数の多少の如何を問わず、文書類似度の計算量が大幅に減少し、ユーザの類似文書検索効率の大幅向上を図ることができる。

【００１１】請求項３記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、検索キー文書を入力する入力手段と、検索対象文書のカテゴリ類似度分布を設定するカテゴリ類似度分布設定手段と、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報情報抽出手段と、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、このカテゴリ類似度算出手段により算出したカテゴリ類似度とカテゴリ類似度分布設定手段により設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜する類似度算出文書選抜手段と、この類似度算出文書選抜手段より選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出する文書類似度算出手段と、この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力する出力手段とを有することを特徴とするものである。

【００１２】請求項４記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、検索キー文書を作成し、検索対象文書のカテゴリ類似度分布を設定し、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜し、選抜した文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に選抜した文書群の各文書類似度を算出し、算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力することを特徴とするものである。

【００１３】請求項３記載の発明に係る類似文書検索装置の構成を使用した請求項４記載の発明に係る類似文書検索方法は、基本的には請求項１、２記載の発明と同様な作用を発揮することに加え、予め検索対象文書のカテゴリ類似度分布を設定しておくことにより、検索キー文書に最も近い文書が含まれているカテゴリを任意選抜し

て検索漏れを防ぎつつユーザの文書検索効率を大幅に向上させることが可能となる作用を発揮する。

【００１４】請求項５記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索装置において、検索キー文書を入力する入力手段と、検索対象文書のカテゴリ類似度分布を設定するカテゴリ類似度分布設定手段と、検索対象文書の文書数分布を設定する文書数分布設定手段と、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出する文書注目文脈情報抽出手段と、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出するカテゴリ注目文脈情報情報抽出手段と、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出するカテゴリ類似度算出手段と、このカテゴリ類似度算出手段により算出したカテゴリ類似度とカテゴリ類似度分布設定手段により設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、前記文書数分布設定手段により設定した検索対象文書の文書数分布に応じた数の文書群に絞る類似度算出文書選抜手段と、この類似度算出文書選抜手段より絞った文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に各文書類似度を算出する文書類似度算出手段と、この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力する出力手段とを有することを特徴とするものである。

【００１５】請求項６記載の発明は、一文書を検索キーとして、文書群が内容別に格納されている文書データベース中から類似文書を抽出する類似文書検索方法において、検索キー文書を作成し、検索対象文書のカテゴリ類似度分布を設定し、検索対象文書の文書数分布を設定し、文書データベース中の各文書及び前記検索キー文書からその内容を示唆する文書注目文脈情報を抽出し、文書データベース中の内容別の文書群を一纏まりとしてカテゴリ注目文脈情報を抽出し、検索キー文書から抽出した文書注目文脈情報と、内容別の一纏まりの文書群から抽出したカテゴリ注目文脈情報とのカテゴリ類似度を算出し、算出したカテゴリ類似度と、設定したカテゴリ類似度分布とを基に、前記文書データベース中から検索キー文書に近似した文書が含まれるカテゴリに属する文書類似度算出用の文書群を選抜するとともに、設定した検索対象文書の文書数分布に応じた数の文書群に絞り、絞った文書群の文書注目文脈情報と検索キー文書の文書注目文脈情報とを基に各文書類似度を算出し、この文書類似度算出手段により算出した各文書類似度に基づき検索した文書の識別情報を検索結果として出力することを特徴とするものである。

【００１６】請求項５記載の発明に係る類似文書検索装置の構成を使用した請求項６記載の発明に係る類似文書検索方法は、基本的には請求項３、４記載の発明と同様な作用を発揮することに加え、文書データベースに格納されている文書数又は指定文書数によって、検索キーと各文書との類似度算出を行う最高文書数を定め、この最高文書数を基準として類似度算出を行う文書に属するカテゴリを定めて文書類似度算出を行う文書数を絞り込んで検索することができ、ユーザの類似文書検索効率を大幅に向上させることができる作用を発揮する。

【００１７】

【発明の実施の形態】以下に、本発明の実施の形態を図面を参照しながら説明する。

【００１８】図１は、本実施の形態の類似文書検索装置の構成を示すブロック図である。本実施の形態の類似文書検索装置は、ＣＰＵ、メモリーから構成される制御装置１と、キーボード等の入力装置２と、類似文書の検索結果等を表示する表示装置３と、検索対象となる文書データ等を格納する外部記憶装置４とを有している。

【００１９】図2は、制御装置１の詳細構成例及び入力装置２、表示装置３を示すブロック図である。

【００２０】制御装置１は制御部とメモリ部から構成されている。

【００２１】制御部は各種制御や処理を実行する部分で、メイン処理部２００、初期化部２０１、入力部２０２、出力部２０３、カテゴリ別文書数設定部２０４、類似度分布設定部２０５、文書数分布設定部２０６、文書データベース注Ｚ目文脈情報抽出部２０７、検索キー文書注目文脈情報抽出部２０８、カテゴリ注目文脈情報情報抽出部２０９、カテゴリ類似度算出部２１０、類似度算出文書選抜部２１１、カテゴリ選抜部２１２、カテゴリ文書選抜部２１３、文書類似度算出部２１４、カテゴリ文書設定部２１５、検索結果出力部２１６等を具備している。

【００２２】メモリ部は、カテゴリ別文書数バッファ部２２０、類似度分布設定バッファ部２２１、文書数分布設定バッファ部２２２、文書データベース注目文脈情報バッファ部２２３、検索キー文書注目文脈情報バッファ部２２４、カテゴリ注目文脈情報バッファ部２２５、カテゴリ類似度バッファ部２２６、類似度算出文書バッファ部２２７、選抜カテゴリバッファ部２２８、選抜文書バッファ部２２９、文書類似度バッファ部２３０、カテゴリ文書バッファ部２３１、作業バッファ部２４０等を具備している。

【００２３】ここで、前記初期化部２０１は、各バッファ部の初期化を行う。入力部２０２は、入力装置２からユーザによって入力される検索キー文書の情報その他各種の情報を作業バッファ部２４０へ出力する。

【００２４】出力部２０３は、入力部２０２により行っ

た検索キー文書や各種設定の内容を表示装置３に出力し表示させる。

【００２５】カテゴリ別文書数設定部２０４は、文書の内容に応じてカテゴリ分けされている文書データベースに対して、その文書データベース中のカテゴリ別の文書数の設定を行う。ここで設定されたカテゴリ別の文書数は、カテゴリ別文書数バッファ部２２０に格納される。

【００２６】類似度分布設定部２０５は、検索キー文書と文書データベース中の各文書における各カテゴリとの類似度値から、その各カテゴリに属する文書群と検索キー文書との類似度算出を行うカテゴリを決定するための設定を行う。

【００２７】この場合の設定では、ユーザの指定した類似度値以上のカテゴリに属する文書群と検索キー文書との類似度算出を行うように設定したり、全カテゴリとの類値度値分布の統計情報（偏差値等）により文書群まで類似度算出を行うようカテゴリを設定したりする。ここで設定された内容は、類似度分布設定バッファ部２２１に格納される。

【００２８】文書数分布設定部２０６は、類似度分布設定部２０５により、検索キー文書と文書データベース中の各カテゴリとの類似度値から、その各カテゴリに属する文書群と検索キー文書との類似度算出まで行うカテゴリを決めるがこれと合わせて、検索キー文書との類似度算出する文書数の最大値を設定する。

【００２９】類似度分布設定部２０５により、各カテゴリに属する文書群と検索キー文書との類似度算出まで行う対象カテゴリが設定されても、そのカテゴリに属する文書数が、文書数分布設定部２０６により設定された値を超えた場合には、文書数分布設定部２０６により設定された値以下の文書数となるように、検索キー文書との類似度が低いカテゴリが、各カテゴリに属する文書群と検索キー文書との類似度算出まで行う対象カテゴリから除外する。ここで設定された内容は、文書数分布設定バッファ部２２２に格納される。

【００３０】文書データベース注目文脈情報抽出部２０７は、文書データベース中に格納されている各文書からその文書の内容を表す上で注目できる文脈情報を抽出する。

【００３１】この場合の文脈情報には、単語種、出現回数、出現位置、共起情報等がある。

【００３２】文脈情報のデータは、各文書単位に作成され、文書データベース注目文脈情報バッファ部２２３に格納される。

【００３３】検索キー文書注目文脈情報抽出部２０８は、入力部２０２から入力される検索キー文書に対して、その文書の内容を表す上で注目できる文脈情報を抽出する。

【００３４】この場合の文脈情報には、単語種、出現回数、出現位置、共起情報等がある。

【００３５】文脈情報のデータは、検索キー文書注目文脈情報バッファ部２２４に格納される。

【００３６】カテゴリ注目文脈情報情報抽出部２０９は、文書データベース注目文脈情報抽出部２０７によって抽出された文脈情報を、文書データベース注目文脈情報バッファ部２２３から呼び出し、また、カテゴリ文書設定部２１５によって作成された文書ＩＤ－カテゴリ情報をカテゴリ文書バッファ部２３１から呼び出し、同カテゴリの各文書の文脈情報を一纏めにし、カテゴリ注目文脈情報バッファ部２２５に格納する。

【００３７】カテゴリ類似度算出部２１０は、検索キー文書注目文脈情報抽出部２０８によって作成された検索キー文書注目文脈情報バッファ部２２４と、カテゴリ注目文脈情報情報抽出部２０９によって作成された文書データベース注目文書情報バッファ部２２３から、検索キー文書と各カテゴリとの類似度を算出する。算出したカテゴリ類似度の値は、カテゴリ類似度バッファ部２２６に格納される。

【００３８】類似度算出文書選抜部２１１は、選抜カテゴリバッファ部２２８に格納されているカテゴリに属する文書ＩＤをカテゴリ文書バッファ部２３１を参照することにより抽出し、類似度算出文書バッファ部２２７から対応する文書ＩＤを選抜して選抜文書バッファ部２２９に格納する。

【００３９】カテゴリ選抜部２１２は、カテゴリ類似度算出部２１０によって算出されたカテゴリ別類似度値をカテゴリ類似度バッファ部２２６から呼び出し、類似度分布設定部２０５によって指定した類似度分布設定バッファ部２２１に格納されている条件に合致するカテゴリを選抜カテゴリバッファ部２２８に格納する。

【００４０】カテゴリ文書選抜部２１３は、カテゴリ文書バッファ部２３１に、検索キー文書と文書データベース中に格納されている各文書との類似度算出を行う文書数が格納されている場合、選抜カテゴリバッファ部２２８に格納されているカテゴリに属する文書数をカテゴリ別文書数バッファ部２２０を参照し類似度算出を行う文書数を算出する。

【００４１】そして、若しその文書数が、選抜文書数バッファ部２２９に格納されている文書数の条件に合致していない場合は、合致するように選抜カテゴリバッファ部２２８に格納されている幾つかのカテゴリを削除する。削除するカテゴリは、カテゴリ類似度バッファ部２２６を参照し、類似度が最も低いカテゴリから順次削除するようにする。

【００４２】文書類似度算出部２１４は、選抜文書バッファ部２２９に格納されている文書ＩＤに対応する文脈情報を文書データベース注目文脈情報バッファ部２２３から抽出し、各文書ＩＤの文脈情報と検索キー文書注目文脈情報バッファ部２２４に格納されている検索キー文書の文脈情報とから類似度を求め、検索キー文書と文書

データベースに格納されている各文書との類似度をそれぞれ文書類似度バッファ部２３０に格納する。

【００４３】カテゴリ文書設定部２１５は、文書データベース中の各文書ファイルに対して一意に決まる文書ＩＤとカテゴリを設定して、カテゴリ文書バッファ部２３１に格納する。

【００４４】検索結果出力部２１６は、文書類似度バッファ部２３０に格納されている各文書類似度値を参照し、最も高い類似度の文書から順に表示装置３に出力する。

【００４５】次に、本実施の形態装置による文書データベース作成手順を図３を参照して、また、類似文書検索手順を図４を参照して各々説明する。

【００４６】まず、文書データベース作成手順について、図３を参照して説明する。

【００４７】まず、初期化部２０１が起動し、メモリ部のクリア等を行う（ステップＳ１０１）。そして、カテゴリ文書設定部２１５が起動し、文書データベースの登録文書に対して、文書ＩＤとカテゴリの設定を行う（ステップＳ１０２）。文書ＩＤは文書データベース中の文書を一意に決めるためのもので重複はない。カテゴリは文書データベース中の各文書を文書の内容ごとに一纏めにするためのもので、文書データベースの文書内容の分類数によりカテゴリ数が決まる。また、１文書に１種類のカテゴリが割り当てられる。文書ＩＤ－カテゴリー文書名（ファイル名）がリンク付けされて、図１６に示すように、カテゴリ文書バッファ部２３１に格納される。

【００４８】次に、カテゴリ別文書数設定部２０４が起動され、文書データベース中のカテゴリ別文書数が設定され（ステップＳ１０３）、カテゴリ別文書数バッファ部２２０に格納される。

【００４９】そして、文書データベース注目文脈情報抽出部２０７が起動し、外部記憶装置４に格納されている文書データベースの各文書からその文書の内容を表す注目文脈情報を作成し、文書データベース注目文脈情報バッファ部２２３に格納する（ステップＳ１０４）。注目文脈情報には、共起情報、例えば、「新製品の発表に関する文書」の場合は、「新製品…発表」が、単語情報としてその文書の内容を表す上で重要な単語がある。

【００５０】文書データベースの各文書について行い、図８に示すように、文書データベース注目文脈情報バッファ部２２３に格納する。

【００５１】続いて、カテゴリ注目文脈情報抽出部２０９が起動し、カテゴリ文書バッファ部２３１と文書データベース注目文脈情報バッファ部２２３を参照し、カテゴリ単位にそのカテゴリに属する文書の注目文脈情報を纏め、カテゴリ注目文脈情報バッファ部２２５に格納する（ステップＳ１０５）。カテゴリ注目文脈情報バッファ部２２５には、図１０に示すように、カテゴリと注目文脈情報がリンク付けられ格納される。これで、文書データベース作成手順をすべて終了する。

【００５２】次に、類似文書検索手順について、図４を参照して説明する。

【００５３】まず、初期化部２０１が起動し、メモリ部のクリア等を行う（ステップＳ２０１）。そして、検索キー文書の入力か、検索実行か、検索設定かを選択する（ステップＳ２０２）。

【００５４】ステップＳ２０２で検索キー文書を選択した場合は、入力部２０２が起動し、入力装置２より図１７に示すような検索キー文書の入力を行い（ステップＳ２０５）、検索キー文書のデータが作業バッファ部２４０に格納される。

【００５５】そして、検索キー文書注目文脈情報抽出部２０８が起動し、作業バッファ部２４０から検索キー文書データを呼び出しその文書の内容を表す注目文脈情報を作成し、図９に示すように、検索キー文書注目文脈情報バッファ部２２４に格納する（ステップＳ２０６）。そして、ステップＳ１０２に戻る。

【００５６】また、ステップＳ２０２で検索設定が選択された場合は、類似度分布設定部２０５が起動し、検索キー文書とカテゴリとの類似度の算出結果から、どのカテゴリに属する文書に対して検索キー文書とのカテゴリ類似度を算出するかを設定する。例えば、図６に示すように検索キー文書と各カテゴリとのカテゴリ類似度がある基準値以上のカテゴリのみに絞りたい場合のその基準値（例えば類似度０．５）を設定する（ステップＳ２０３）。又は、検索キー文書と各カテゴリとのカテゴリ類似度の統計をとり、標準偏差等から設定することもできる。類似度分布の設定内容は図６に示すように類似度分布設定バッファ部２２１に格納される。

【００５７】次に、文書数分布設定部２０６が起動し、検索キー文書と各カテゴリとのカテゴリ類似度からの絞り込みでも文書数が期待値以下にすることができなかった場合のために、検索キー文書との文書類似度を算出する文書数（例えば４０００以下）を設定する（ステップＳ２０４）。

【００５８】設定した文書数のデータは図７に示すように文書数分布設定バッファ部２２２に格納される。そして、ステップＳ１０２に戻る。

【００５９】また、ステップＳ２０２で検索実行が選択された場合は、カテゴリ類似度算出部２１０が起動し、カテゴリ注目文脈情報バッファ部２２５と検索キー文書注目文脈情報バッファ部２２４を参照し、検索キー文書と各カテゴリとのカテゴリ類似度を算出し（ステップＳ２０７）、図１１に示すように、カテゴリ類似度バッファ部２２６に格納する。

【００６０】そして、カテゴリ選抜部２１２が起動し、類似度分布設定バッファ部２２１に類似度分布が設定されているか否かを判断し（ステップＳ２０８）、もし、ステップＳ２０８において、類似度分布設定バッファ部

２２１に類似度分布が設定されている場合は、類似度分布設定バッファ部２２１に格納されている条件に合致したカテゴリ類似度のカテゴリ（例えば、２、４等）を、図１３に示すように、選抜カテゴリバッファ部２２８に格納する（ステップＳ２０９）。

【００６１】つぎに、カテゴリ文書選抜部２１３が起動し、文書数分布設定バッファ部２２２に文書数分布のデータが設定されている場合は（ステップＳ２１０）、選抜カテゴリバッファ部２２８とカテゴリ別文書数バッファ部２２０を呼び出し、選抜カテゴリバッファ部２２８に格納されてる各カテゴリに属する文書数の合計が、文書数分布設定バッファ部２２２に格納される条件に合致していない場合は、カテゴリ類似度バッファ部２２６を呼び出し、選抜カテゴリバッファ部２２８に格納されている最も低い類似度のカテゴリを選抜カテゴリバッファ部２２８から順次削除し、文書数分布設定バッファ部２２２に格納される条件に合致する文書数となるようにする（ステップＳ２１１）。

【００６２】ステップＳ２０８において、類似度分布設定バッファ部２２１に類似度分布が設定されていない場合は、カテゴリ選抜部２１２が起動し、カテゴリ類似度バッファ部２２６において最も高いカテゴリ類似度のカテゴリを選抜して（ステップＳ２１２）、選抜カテゴリバッファ部２２８に格納する。

【００６３】次に、類似度算出文書選抜部２１１が起動し、選抜カテゴリバッファ部２２８に格納されているカテゴリに対応する文書ＩＤをカテゴリ文書バッファ部２３１を参照して図１３に示す選抜文書バッファ部２２９に格納する（ステップＳ２１３）。

【００６４】次に、文書類似度算出部２１４が起動し、選抜文書バッファ部２２９に格納されている各文書ＩＤについてその文書に対応する注目文脈情報を文書データベース注目文脈情報バッファ部２２３から呼び出し、検索キー文書注目文脈情報バッファ部２２４との文書類似度を算出し（ステップＳ２１４）、その結果を文書類似度バッファ部２３０に図１５に示すように「文書ＩＤ－類似度」として対応付けて格納する。

【００６５】次に、選択文書バッファ部２２９に格納されている全ての文書ＩＤの文書と検索キー文書との文書類似度を算出し、各文書ＩＤごとにその文書類似度を文書類似度バッファ部２３０に格納する（ステップＳ２１４）。

【００６６】そして、検索結果出力部２１６が起動し、文書類似度バッファ部２３０に格納されている高類似度の文書を結検索処理結果として、図１８に示すように、出力部２０３から出力装置３に出力する（ステップＳ２１５）。

【００６７】最後に、類似検索の処理を続けるか否か判断し（ステップＳ２１６）、処理を継続する場合はステップＳ２０２に移行し、処理を終了する場合は全ての検索処理の終了となる。

【００６８】なお、本発明は上記の実施の形態に限定されるものではない。

【００６９】

【発明の効果】請求項１及び２記載の発明によれば、文書数の多少の如何を問わず、文書類似度の計算量が大幅に減少し、ユーザの類似文書検索効率の大幅向上を図ることができる類似文書検索装置及びこの類似文書検索装置を用いた類似文書検索方法を提供できる。

【００７０】請求項３及び４記載の発明によれば、請求項１、２記載の発明と同様な効果を奏することに加え、検索キー文書に最も近い文書が含まれているカテゴリを任意選抜して検索漏れを防ぎつつ類似文書検索効率の大幅向上を図ることができる類似文書検索装置及びこの類似文書検索装置を用いた類似文書検索方法を提供できる。

【００７１】請求項５及び６記載の発明によれば、請求項３、４記載の発明と同様な効果を奏することに加え、文書類似度算出を行う文書数を絞り込んで検索することで類似文書検索効率の大幅向上を図ることができる類似文書検索装置及びこの類似文書検索装置を用いた類似文書検索方法を提供できる。

【図面の簡単な説明】

【図１】本発明の実施の形態装置の概略構成を示すブロック図である。

【図２】本発明の実施の形態装置の制御部のブロック図である。

【図３】本実施の形態の文書データベース作成の手順を示すフロチャートである。

【図４】本実施の形態の類似文書検索の手順を示すフロチャートである。

【図５】本実施の形態のカテゴリ別文書数バッファ部格納例を示す図である。

【図６】本実施の形態の類似度分布設定バッファ部格納例を示す図である。

【図７】本実施の形態の文書数分布設定バッファ部格納例を示す図である。

【図８】本実施の形態の文書データベース注目文脈情報バッファ部格納例を示す図である。

【図９】本実施の形態の検索キー文書注目文脈情報バッファ部格納例を示す図である。

【図１０】本実施の形態のカテゴリ注目文脈情報バッファ部格納例を示す図である。

【図１１】本実施の形態のカテゴリ類似度バッファ部格納例を示す図である。

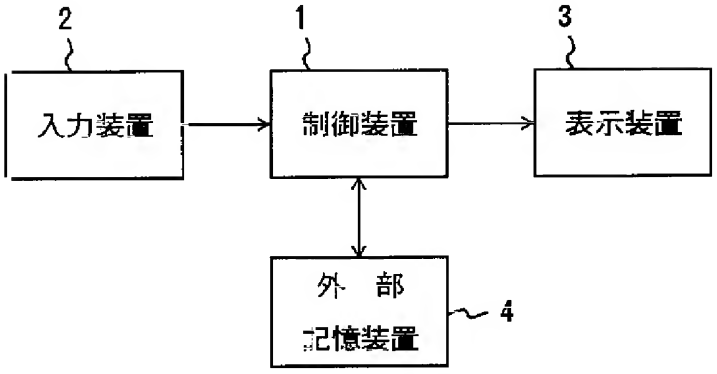
【図１２】本実施の形態の類似度算出文書バッファ部格納例を示す図である。

【図１３】本実施の形態の選抜カテゴリバッファ部格納例を示す図である。

【図１４】本実施の形態の選抜文書バッファ部格納例を

示す図である。
【図１５】本実施の形態の文書類似度バッファ部格納例を示す図である。
【図１６】本実施の形態のカテゴリ文書バッファ部格納例を示す図である。
【図１７】本実施の形態の検索キー文書の入力例を示す図である。
【図１８】本実施の形態の検索結果の出力例を示す図である。
【符号の説明】
１　制御装置
２　入力装置
３　表示装置
４　外部記憶装置
２００　メイン処理部
２０１　初期化部
２０２　入力部
２０３　出力部
２０４　カテゴリ別文書数設定部
２０５　類似度分布設定部
２０６　文書数分布設定部
２０７　文書データベース注目文脈情報抽出部
２０８　検索キー文書注目文脈情報抽出部
２０９　カテゴリ注目文脈情報情報抽出部
２１０　カテゴリ類似度算出部
２１１　類似度算出文書選抜部
２１２　カテゴリ選抜部
２１３　カテゴリ文書選抜部
２１４　文書類似度算出部
２１５　カテゴリ文書設定部
２１６　検索結果出力部
２２０　カテゴリ別文書数バッファ部
２２１　類似度分布設定バッファ部
２２２　文書数分布設定バッファ部
２２３　文書データベース注目文脈情報バッファ部
２２４　検索キー文書注目文脈情報バッファ部
２２５　カテゴリ注目文脈情報バッファ部
２２６　カテゴリ類似度バッファ部
２２７　類似度算出文書バッファ部
２２８　選抜カテゴリバッファ部
２２９　選抜文書バッファ部
２３０　文書類似度バッファ部
２３１　カテゴリ文書バッファ部
２４０　作業バッファ部

【図１】



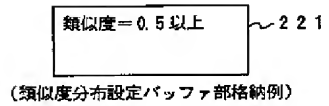
【図５】

カテゴリ	文書数
1	2151
2	2155
3	1113
4	1555
5	1222
…	…

220

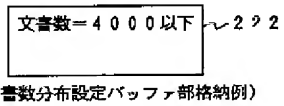
（カテゴリ別文書数バッファ部格納例）

【図６】



（類似度分布設定バッファ部格納例）

【図７】



（文書数分布設定バッファ部格納例）

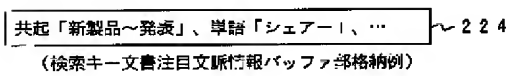
【図８】

223

文書ＩＤ	注目文脈情報
1	共起「新製品～発表」２回、単語「20%シェアー」３回、…
2	共起「記念誌～発行」、単語「市場動向」「東京」
3	共起「設計報告書～作成」「パソコン製造」
4	共起「会議開催～通知」、「委員会」
…	…

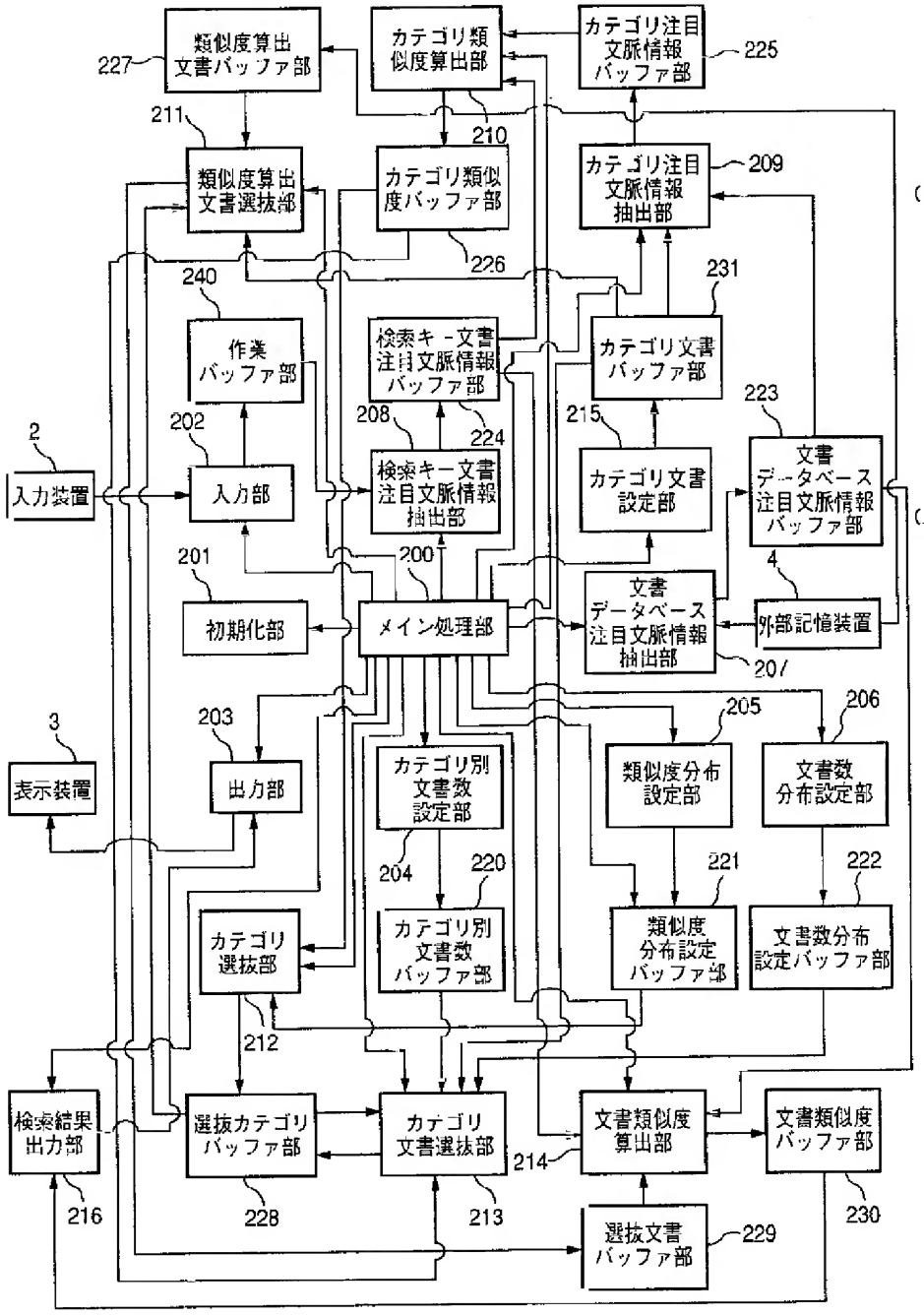
（文書データベース注目文脈情報バッファ部格納例）

【図９】



（検索キー文書注目文脈情報バッファ部格納例）

【図２】



【図１１】

カテゴリ	類似度
1	0.5555
2	0.2500
3	0.6211
…	…

226

（カテゴリ類似度バッファ部格納例）

【図１５】

文書ＩＤ	類似度
３４	0.98
５４	0.23
７７	0.54
…	…

230

（文書類似度バッファ部格納例）

【図１２】

３４、５４、６７、７７、７８、７９、９０、１２３、…

227

（類似度算出文書バッファ部格納例）

【図１３】

２、４

228

（選抜カテゴリバッファ部格納例）

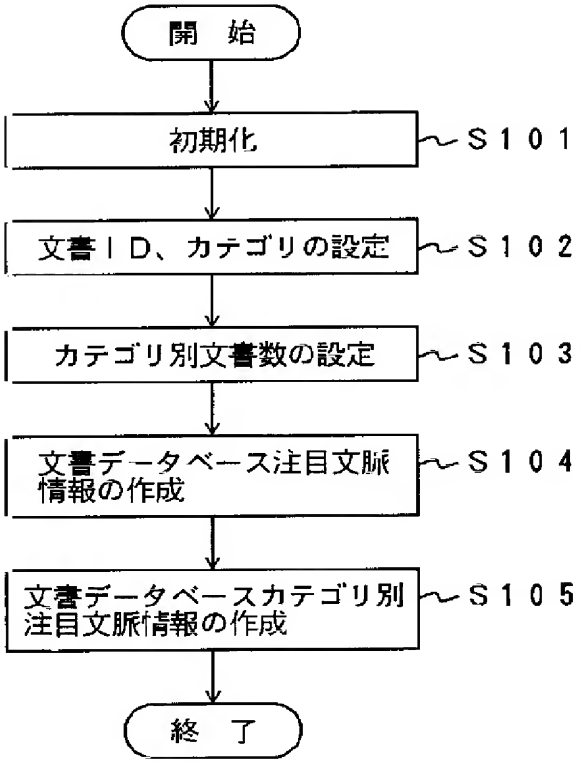
【図１４】

３４、５４、７７、７８、９０、１２３、…

229

（選抜文書バッファ部格納例）

【図３】



【図１０】

225

カテゴリ	注目文脈情報
1	共起「新製品～発表」「新商品～販売」単語「シェアー」「市場」
2	共起「会議開催～通知」「議事録～配布」「委員会」「会議」
…	

（カテゴリ注目文脈情報バッファ部格納例）

【図１８】

検索結果：
1．文書ＩＤ＝３４　　類似度＝０．９８　～３
2．文書ＩＤ＝７７　　類似度＝０．５２

（検索結果の出力例）

【図１６】

231

文書ＩＤ	カテゴリ	ファイル名
1	3	/usr/date/text1.txt
2	5	/usr/data/text2.txt
3	2	/usr/date/text3.txt
…	…	…

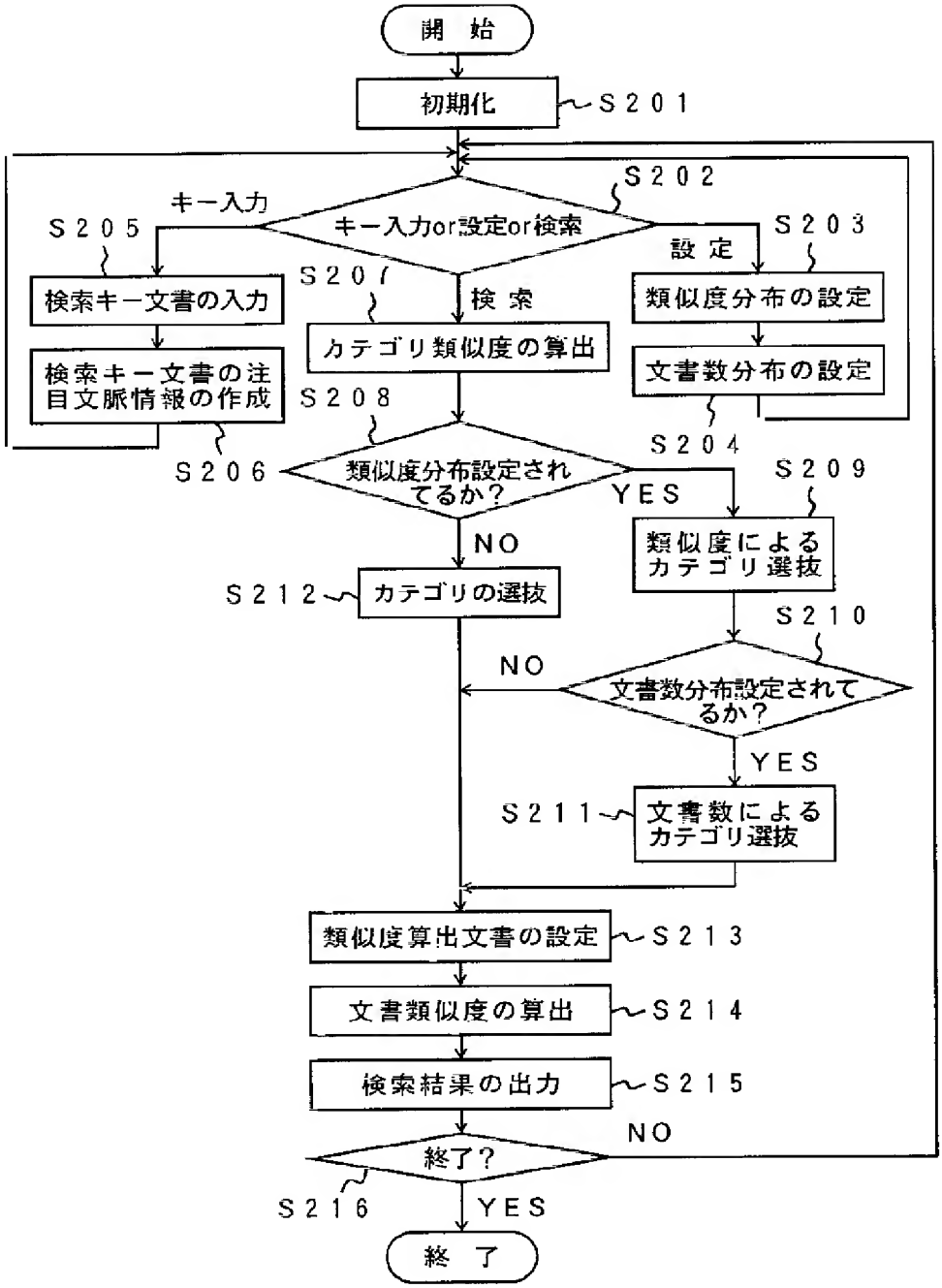
（カテゴリ文書バッファ部格納例）

【図１７】

検索キー文書の入力＝/usr/date/key.txt

（検索キー文書の入力例）

【図４】



フロントページの続き

(72)発明者　中本　幸夫
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　仁科　卓哉
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　久保田　直秀
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内